한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 30, No. 4, P. 19~29

의료영상 분류를 위한 심층신경망 훈련에서 StyleGAN 합성 영상의 데이터 증강 효과 분석

이한상[1○] 우아라[2○] 홍헬렌[2*]

[1]한국과학기술원 정보전자연구소
[2]서울여자대학교 소프트웨어융합학과
○ 공동 1저자
* 교신저자

hansanglee@kaist.ac.kr
{war00, hlhong}@swu.ac.kr

Data Augmentation Effect of StyleGAN-Generated Images in Deep Neural Network Training for Medical Image Classification

Hansang Lee[1○] Arha Woo[2○] Helen Hong[2*]

[1]School of Electrical Engineering, Information & Electronics Research Institute, KAIST
[2]Department of Software Convergence, Seoul Women's University
○ Co-first authors
* Corresponding author

요 약

본 논문에서는 의료 영상 분류를 위한 심층 신경망 훈련에서 StyleGAN 합성 영상의 데이터 증강 효과를 분석한다. 이를 위해 흉부 X선 영상에서의 폐렴 진단과 복부 CT 영상에서의 간전이암 분류 문제에서 StyleGAN 합성 영상을 이용하여 VGG-16 심층 합성곱 신경망 훈련을 수행한다. 실험에서 분류 결과에 대한 정량적, 정성적 분석을 통해 StyleGAN 데이터 증강이 특징 공간에서 클래스 외곽을 확장하는 특성을 보이며, 이와 같은 특성으로 인해 실제 영상과의 적절한 비율을 통해 혼합했을 때 분류 성능이 개선될 수 있음을 확인하였다.

Abstract

In this paper, we examine the effectiveness of StyleGAN-generated images for data augmentation in training deep neural networks for medical image classification. We apply StyleGAN data augmentation to train VGG-16 networks for pneumonia diagnosis from chest X-ray images and focal liver lesion classification from abdominal CT images. Through quantitative and qualitative analyses, our experiments reveal that StyleGAN data augmentation expands the outer class boundaries in the feature space. Thanks to this expansion characteristics, the StyleGAN data augmentation can enhance classification performance when properly combined with real training images.

키워드: 심층신경망, 의료영상, 영상 분류, 데이터 증강, 생성적 적대 신경망
Keywords: Deep learning, Medical imaging, Image classification, Data augmentation, Generative adversarial network

*corresponding author: Helen Hong/Seoul Women's University(hlhong@swu.ac.kr)

Received : 2024.06.08./ Review completed : 1st 2024.08.02. / Accepted : 2024.08.13.
DOI : 10.15701/kcgs.2024.30.4.19
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

데이터 증강(data augmentation)은 심층신경망 등의 기계학습에서 주어진 데이터에 인위적인 변화를 가해 데이터의 패턴과 양을 증가시키는 과정이다. 데이터 증강은 기계학습의 효율을 개선하고 과적합(overfitting)을 방지하는 등 기계학습에서 필수적인 과정으로 알려져 있다. 특히 의료 영상 분류 문제에서 의료 영상 데이터는 종종 샘플 크기가 작고, 클래스 불균형과 같은 제약을 가지고 있다 [1, 2]. 이러한 제약은 의료 영상 분류에서 심층신경망의 학습 효율을 저해하기 때문에 데이터 증강은 인위적으로 훈련 데이터셋을 확장하여 이러한 문제들을 해결하는 데 있어 중요한 역할을 한다. 의료 영상 분류에서 데이터 증강은 훈련 데이터의 양을 증가시키고 패턴을 다양화시키기 위한 다양한 기술들을 포함하는데, 일반적인 방법으로는 확대 및 축소, 수직 및 수평 이동, 회전, 뒤집기 등의 어파인 변환이 있다.

최근 의료 영상에서 활발히 사용되는 데이터 증강 기법 중 하나는 생성적 적대 신경망(Generative Adversarial Network, GAN) 학습을 통한 영상 생성 기법이다. GAN은 실제 영상과 구별하기 어려운 사실적인 합성 영상을 생성할 수 있는 심층신경망 기술로, 최근에는 DCGAN(Deep Convolutional GAN) [3], PGGAN(Progressive Growing of GANs) [4], StyleGAN(Style-Based GAN) [5] 등 다양한 개선 모델들이 제안되어 영상 생성 및 데이터 증강 관련 작업에 활용되고 있다. 의료 영상에서 GAN 기반 데이터 증강을 통해 분류 성능을 향상시킨 연구로는 국소 간 병변을 분류하는 문제에서 DCGAN 기법을 사용하여 데이터를 증강하고 AlexNet을 훈련시켜 분류 성능을 향상시킨 연구 [6], 국소 간 병변을 분류하는 문제에서 StyleGAN과 MixUp, AugMix 데이터 증강을 통해 VGG-16 네트워크의 분류성능 향상 효과를 비교한 연구 [7], 국소 간 병변을 분류하는 문제에서 F&BGAN(Forward and Backward GAN)을 제안하고 M-VGG-16(Multi-scale VGG-16) 네트워크를 훈련시켜 분류 성능을 향상시킨 연구 [8], 국소 간 병변을 분류하는 문제에서 어파인 변환 기반 데이터 증강 기법과 DCGAN을 사용하여 영상 데이터를 보강하고 mask-to-image 변환을 통해 보강된 영상 데이터에 대응되는 병변 분할 마스크(lesion segmentation mask)를 생성하여 작은 병변에 대한 합성곱 신경망(Convolutional Neural Network; CNN)의 학습 효율성을 위해 LINA 패치를 제안한 연구 [9], 국소 간 병변을 분류하는 문제에서 어파인 변환 기반 데이터 증강 기법과 StyleGAN을 이용하여 의료 영상 데이터를 보강하고 VGG-16 네트워크를 훈련시켜 StyleGAN의 효과를 분석한 연구 [10] 등이 수행되었다. 또한, 흉부 X선 영상에서 분류 성능을 향상시킨 연구로는 흉부 X선 영상을 COVID-19, 정상, 세균성 폐렴(pneumonia bacterial), 바이러스성 폐렴(pneumonia virus)으로 다중 분류하는 문제에서 GAN과 심층 전이 학습(deep transfer learning)을 결합한 모델을 제안하고, Googlenet, ResNet18, AlexNet을 훈련시켜 데이터를 보강하고 바이러스 감지의 정확도를 높인 연구 [11], 흉부 X선 영상을 폐렴, 정상, COVID-19으로 다중 분류하는 문제에서 IAGAN(Inception-Augmentation GAN)을 제안하고 U-Net을 훈련시켜 데이터를 효과적으로 보강하고 질병의 분류 정확도를 향상시킨 연구 [2], 흉부 X선 영상을 폐렴, 정상, COVID-19로 분류하는 문제에서 UNET과 CGAN을 결합한 HCUGAN(Hybrid Cyclic UNET GAN)을 제안하여 데이터셋의 크기를 향상시킨 연구 [12] 등이 수행되었다.

본 논문에서는 의료 영상 분류를 위한 합성곱 신경망 학습에 있어 StyleGAN 합성 영상의 데이터 증강 효과를 평가 및 검증하고자 한다. 의료영상 분류에 있어 GAN 기반 데이터 증강을 통해 분류 효율을 개선했음을 정량적으로 보고한 연구는 다수 수행되었으나, GAN 합성 영상이 분류 효율을 개선하는 데 어떤 역할을 하는지에 대한 정성적, 정량적 관점에서 분석하고 검증한 연구는 아직 보고되지 않았다. 본 연구에서는 흉부 X선 영상에서의 폐렴 진단 문제와 복부 CT 영상에서의 간전이암 분류 문제의 두 가지 의료 영상 분류 문제에 대해 StyleGAN 데이터 증강을 통한 신경망 학습을 적용하여 데이터셋 편향에 강인한 분석을 수행한다. 각각의 데이터셋에서 StyleGAN 데이터 증강의 효과를 분석하기 위해 첫째로 StyleGAN 합성 영상에 대한 육안 평가를 통해 StyleGAN 영상 생성의 성능과 한계를 분석한다. 둘째로 StyleGAN 합성 영상이 포함된 데이터로 훈련된 분류기에 대한 분류 정확도(accuracy), F1 score, 정밀도(precision) 및 재현율(recall) 수치 분석을 통한 정량적 평가를 수행한다. 마지막으로 훈련 데이터에서의 StyleGAN 합성 영상의 혼합 비율에 따른 분류기 특징값 공간(feature space)의 클래스 별 데이터 분포를 tSNE(t-distributed stochastic neighbor embedding) [13] 시각화를 통해 분석함으로써 StyleGAN 합성 영상이 클래스 분포를 정의하는 데에 어떤 역할을 하는지, StyleGAN 합성 영상과 실제 영상의 혼합 비율이 분류기의 분류 효율에 어떻게 영향을 미치는지를 정성적으로 분석한다.

2. 제안 방법

그림 1은 제안한 방법의 개요도를 나타낸다. 첫째, 전처리를 거친 의료영상에 대해 어파인 변환 기반 데이터 증강 및 StyleGAN 영상 생성을 통해 각각 데이터를 증강한다. 어파인 변환을 통해 생성한 실제 영상(real image)과 StyleGAN을 통해 생성한 합성 영상(synthetic image)을 혼합하여 훈련 데이터를 구성한다. 둘째, 훈련 데이터에 대해 합성곱 신경망 분류 네트워크를 학습시키고 실험 데이터 분류 및 평가를 수행한다.

2.1 데이터 전처리

데이터 증강 및 훈련에 앞서 영상 크기 조정 및 밝기값 정규화를 통한 데이터 전처리를 수행한다. 흉부 X선 데이터는 영상 크기 및 해상도가 다양하게 분포해 있으나 데이터 전처리 단계에서 모든 스캔의 크기를 128×128 픽셀로 정규화하고, 색상 값의 범위를 0 1 사이로 조정한 뒤 평균 0.5, 표준 편차 0.5로 밝기값 정규화를

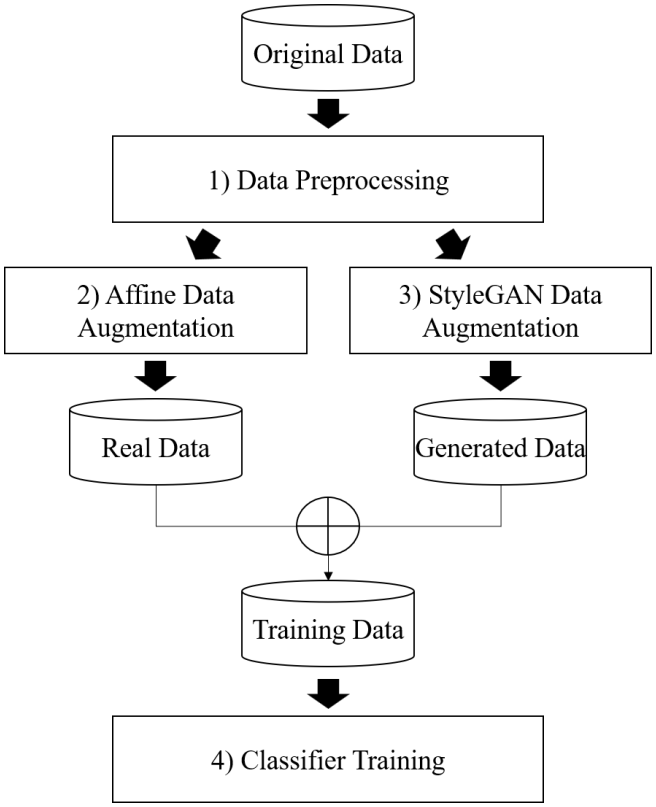


Figure 1: Overview of the proposed method

수행한다. 복부 CT 데이터는 입력 영상의 크기를 512 × 512 영상 크기에서 수동분할된 병변을 중심으로 64×64 픽셀 크기의 패치 영상을 추출하고, 색상 값의 범위를 0 1 사이로 밝기값 정규화를 수행한다.

2.2 어파인 변환 기반 데이터 증강

일반적으로 의료 영상 분류 및 분할 문제 등에서는 데이터 증강을 위해 어파인 변환 기반 데이터 증강 기법을 사용하는데, 이동, 회전(rotation), 수평 뒤집기(horizontal flipping), 자르기, 확대 및 축소 등이 있다 [14]. 이러한 어파인 변환 기반 데이터 증강은 원본 데이터의 공간적 정보를 변형하기 때문에 생성된 데이터는 기하학적 구조가 유지되며 원본 데이터의 속성과 유사한 특징을 가지는 특징이 있다. 그러나 이러한 데이터 증강은 모델이 데이터의 다양한 속성 변화를 학습하지 못하여 모델의 일반화 성능 향상에 큰 영향을 주지 않는 경우가 발생할 수 있다 [15].

2.3 StyleGAN 기반 데이터 증강

GAN은 생성 모델 중 하나로 영상을 생성하는 생성기(generator)와 합성 영상을 분류하는 판별기(discriminator) 네트워크가 적대적으로 학습하는데, 생성기는 랜덤 잡음을 초기값으로 입력 받아 영상을 생성하고, 판별기는 입력 영상의 특징들을 학습하여 가짜와 진짜를 0과 1로 판별한 뒤 생성기 네트워크를 업데이트하는 과정을 반복하여 실제 데이터의 분포에 가까운 새로운 영상을

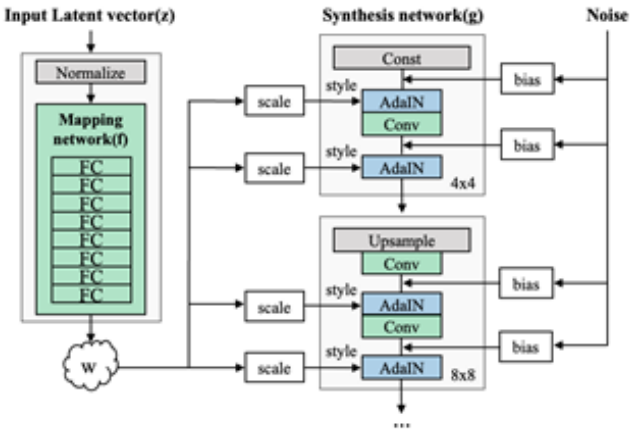


Figure 2: Architecture of the StyleGAN synthesis network.

생성한다. 하지만 이러한 기존의 GAN 구조에서는 스타일이나 특징을 제어하기 어렵기 때문에 본 연구에서는 Wasserstein loss를 손실 함수로 사용하고 저해상도부터 고해상도까지 점진적으로 영상을 생성하는 PGGAN 구조에 스타일 전이(style transfer) 개념을 적용하여 고해상도 이미지를 생성할 수 있고 스타일과 특징을 조절할 수 있는 StyleGAN을 사용한다.

그림 2는 StyleGAN의 네트워크 구조를 나타낸 것으로 입력 벡터 z로부터 직접 영상을 생성하는 대신 매핑 네트워크를 통해 중간 잠재 벡터(intermediate latent vector) w로 변환하여 스타일이 잘 분리될 수 있도록 한다. 이때, w는 합성망(synthesis network)이 영상을 생성하는 과정에서 여러 scale에 다양한 스타일을 넣기 위해 사용되며, 영상의 확률적 다양성(stochastic variation)을 조절하기 위한 잡음과 함께 각 합성곱 계층 이후의 AdaIN(Adaptive Instance Normalization) 연산에 사용된다. AdaIN 연산은 각 층마다 정규화를 수행하고 스타일을 입히는 과정을 통해 최종 출력 영상이 생성된다.

이와 같이 StyleGAN은 매핑 네트워크를 통해 선형성을 가지는 분포를 형성하여 특징을 분리하기 용이해지고, w는 데이터의 분포에 따라 샘플링 할 필요가 없게 되어 분리된 표현을 기반으로 보다 현실적인 영상 생성이 가능하다는 장점이 있다. 이를 통해 StyleGAN은 기하학적 구조와 데이터의 분포를 유지하면서 원본 데이터의 속성과 유사한 새로운 영상을 생성하는 데에 효과적이다.

2.4 분류기 학습 및 평가

데이터 증강을 통해 생성한 실제 영상(real image) 및 합성 영상(synthetic image)을 혼합하여 증강 훈련 데이터(augmented training data)를 구성한 뒤 신경망 분류기 학습 및 평가를 수행한다. 흉부 X선 영상 분류에서 사용된 신경망 모델은 그림 3의 구조도와 같이 5개의 합성곱 계층(convolution layer)과 배치 정규화(batch normalization) 계층, 드롭아웃(dropout) 계층, 최대 풀링(max pooling) 계층 등으로 구성된다. 150×150 픽셀 크기의 영

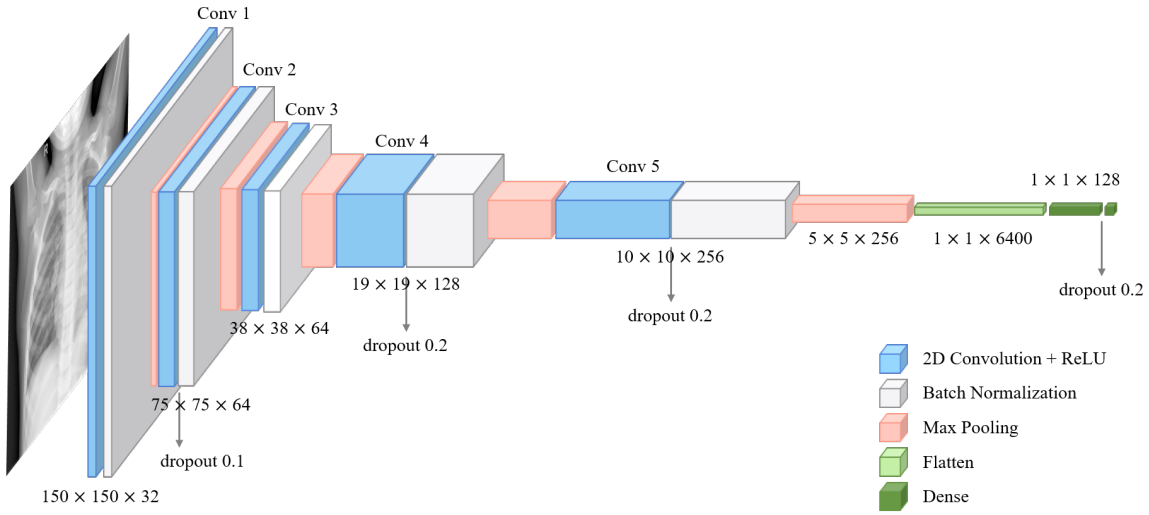


Figure 3: Architecture of Chest X-ray Pneumonia Classifier.

상을 입력받아 32개의 필터를 가진 합성곱 계층 및 배치 정규화 계층, 최대 풀링 계층을 거친다. 이어서 64개의 필터를 가진 합성곱 계층을 지나 드롭아웃 계층과 배치 정규화 계층을 적용하고, 최대 풀링 계층을 거치는 과정을 반복한다. 모든 합성곱 계층에서 3×3 크기의 필터를 사용하고 필터의 개수는 32에서 256까지 점진적으로 증가하며 각 합성곱 계층의 활성화 함수는 ReLU를 사용한다. 마지막 특징값 맵(feature map)은 평탄화(flatten)되며 128-유닛 완전 연결 계층을 적용하고, 마지막 층은 이진 분류를 위해 시그모이드(sigmoid) 활성화 함수를 사용한 1-유닛 완전 연결 계층으로 구성된다. RMSprop 옵티마이저와 이진 크로스 엔트로피 손실 함수를 사용하여 학습을 수행한다.

복부 CT 영상 분류에서 사용된 신경망 모델은 ImageNet 데이터 셋에서 사전 학습된 VGG-16 네트워크 기반의 변형된 모델 [16]을 사용한다. 64×64 크기 영상을 입력받는 이 신경망은 13개의 합성곱 계층과 3개의 완전 연결 계층으로 이루어져 있으며, 모든 합성곱 계층에서 3×3 필터를 사용한다. 본 연구에서는 VGG-16의 합성곱 계층 뒤에 전역 평균 풀링, 완전 연결 계층과 드롭아웃을 연결한다. 마지막 분류기 직전단의 완전 연결 계층의 노드 수는 4096에서 416으로 변경한다. 낭종, 혈관종, 전이암 3개의 클래스로 분류하기 위하여 드롭아웃 계층 뒤에 3-유닛 완전 연결 계층을 추가하고, 마지막 연결 계층에 소프트맥스(softmax) 함수를 적용한다. 활성화 함수는 ReLU 함수를 사용하며, VGG-16 합성곱 계층 앞 단의 55%를 동결하여 전이 학습을 진행한다.

3. 실험 및 결과

3.1 실험 데이터 및 환경

폐렴 진단을 위한 흉부 X선 영상 데이터는 Kaggle 공개 데이터셋을 사용하였다 [1]. 이 데이터는 중국 광저우 여성아동병원(Guangzhou Women and Children's Medical Center)에서 1–5세의 소아 환자들의 후향적 연구로부터 선택된 5856개의 흉부 X선 영상(전방-후방) 데이터로 폐렴(pneumonia)과 정상(normal)의 두 클래스로 구성되어 있으며, 각 클래스에 대한 환자수의 구성은 표 1와 같다. 정상과 폐렴환자 모두 1000명 이상의 훈련데이터가 수집되어 있으나 정상과 폐렴환자 간 비율이 1:3에 가까운 클래스 불균형(class imbalance)을 보인다. 모든 데이터는 환자의 일상적인 임상 관리 일환으로 수행되었고, 기관생명윤리위원회(IRB)의 승인을 받았다. 흉부 X-선 영상 분석을 위해 먼저, 품질 관리를 위해 저품질이거나 판독 불가능한 스캔을 제거하여 스크리닝되었고, 두 명의 전문의에 의해 평가된 후 AI 시스템 훈련을 위해 승인되었으며, 평가 데이터셋에서는 등급 부정확성을 고려하기 위해 세 번째 전문가에 의해 확인되었다 [17].

간전이암 진단을 위한 복부 CT 영상 데이터는 세브란스병원에서 기관생명윤리위원회(IRB)의 승인을 받아 2005년 1월부터 2010년 12월 사이에 502명의 대장암 환자로부터 획득한 1290개의 복부 CT 영상 [9]을 사용하였다. 데이터는 낭종(cyst), 혈

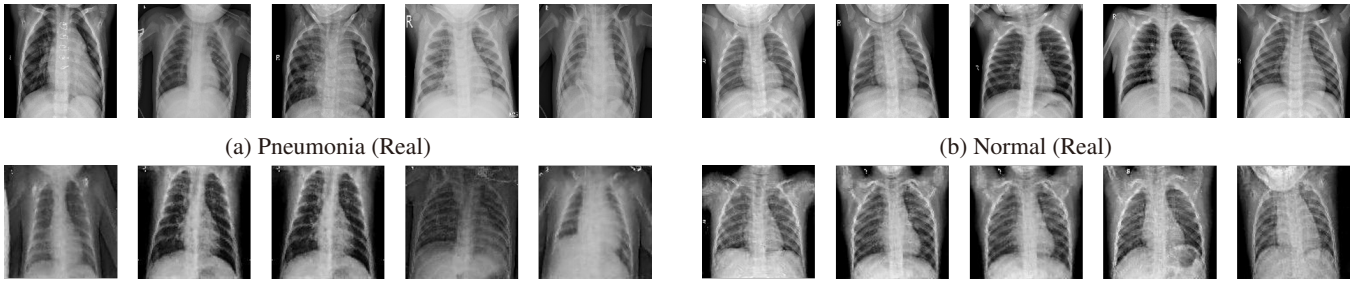
[1] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia?datasetId=17810

Table 1: Statistics of chest X-ray pneumonia classification dataset

Subsets	Normal	Pneumonia	Total
Training	1341	3875	5216
Validation	8	8	16
Test	234	390	624
Total	1583	4273	5856

Table 2: Statistics of abdominal CT FLL classification dataset.

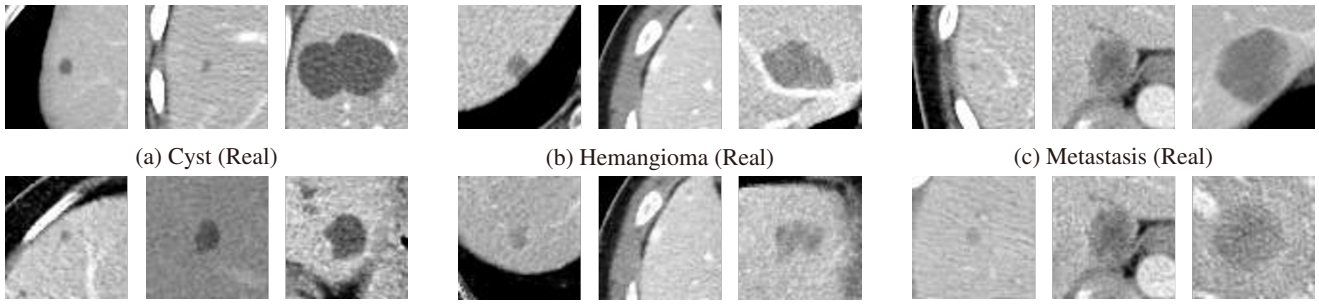
Subsets	Cyst	Hemangioma	Metastasis	Total
Training	433	70	178	681
Validation	115	30	157	302
Test	128	30	149	307
Total	676	130	484	1290



(a) Pneumonia (Real) (b) Normal (Real)

(c) Pneumonia (StyleGAN) (d) Normal (StyleGAN)

Figure 4: Examples of real and StyleGAN-generated chest X-ray images.



(a) Cyst (Real) (b) Hemangioma (Real) (c) Metastasis (Real)

(d) Cyst (StyleGAN) (e) Hemangioma (StyleGAN) (f) Metastasis (StyleGAN)

Figure 5: Example images of real and StyleGAN-generated focal liver lesion CT images.

관종(hemangioma), 전이암(metastasis)의 세 클래스로 구성되어 있으며 각 클래스에 대한 환자수의 구성은 표 2와 같다. 흉부 X선 데이터에 비해 훈련 데이터의 총 환자수가 1000명이 안 되는 소규모 데이터이며, 가장 작은 클래스인 혈관종의 환자수와 가장 큰 클래스인 낭종의 환자수의 비가 1:6에 가까운 심한 클래스 불균형을 보인다. 모든 데이터는 단일 기관의 다중 검출 CT(multidetector CT; MDCT)를 통해 수집되었다. 모든 스캔의 절편 두께(slice thickness)는 3–5mm이고, 해상도는 512×512 픽셀이며, 픽셀 크기는 $0.5\times0.5\text{mm}^2$ – $0.8\times0.8\text{mm}^2$이다.

실험은 양 데이터셋 모두 동일하게 StyleGAN 데이터 증강은 주피터 노트북(jupyter notebook) 6.4.3 환경에서 파이썬(python) 3.6.13 버전을 이용하여 실행하였으며, 텐서플로우(tensorflow) 1.14.0 버전과 파이토치(pytorch) 1.5.1 버전 라이브러리를 이용하였다. 또한 합성곱 신경망 분류 실험은 코랩(colab) 환경에서 파이썬 3.10.12 버전을 이용하여 실행하였으며, 텐서플로우 2.12.0 버전과 파이토치 2.0.1+cu118 버전 라이브러리를 이용하였다.

3.2 실험 계획 및 방법

실험에서는 두 데이터셋 모두에 대해 어파인 변환 기반 데이터 증강 기법과 StyleGAN을 적용하고 학습시킨 뒤 분류 결과를 비교했다. 어파인 변환 기반 데이터 증강 기법으로는 20% 확대 및 축소, ±10% 수평 이동, ±10% 수직 이동, 수평 뒤집기를 적용하였다. 흉부 X선 영상에서 StyleGAN 데이터 증강 하이퍼파라미터는 두 클래스 모두 배치 사이즈(batch size) 32, 학습률(learning rate) 0.001로 설정하였고, 정상의 반복(itetation)은 70000, 폐렴의 반복은 160000으로 설정하였다. 복부 CT 영상에서 StyleGAN 데이터 증강 하이퍼파라미터는 세 클래스 모두 배치 사이즈 8, 학습률 0.001, 반복 70000으로 설정하였다. StyleGAN 영상의 수와 클래스 구성은 실제 훈련 영상과 동일하게 생성하였다. 실험에서는 각 미니배치(mini batch) 내부에서 StyleGAN 합성 영상의 랜덤 샘플의 개수와 실제 영상의 랜덤 샘플의 개수를 조절하는 방식으로 StyleGAN 합성 영상과 실제영상 간의 비율을 조절하였다. 비율은 전체 훈련 데이터에서 StyleGAN 합성 영상이 차지하는 비율로 정의하였으며, 0%, 25%, 50%, 75%, 100%의 총 5가지 비율을 적용하였다. 어파인 변환 기반 데이터 증강과 StyleGAN을 적용한 분류기의 하이퍼파라미터는 에폭(epoch) 12, 배치 사이즈 32, 학습률 0.001을 사용하였고, 드롭아웃 계층의 확률은 0.1로 지정하였으며, 학습률 감소(ReduceLROnPlateau) 콜백(callback) 함수를 적용하여 검증 정확도(val_accuracy)를 모니터링하며 2번의 인내(patience) 기간 동안 정확도의 향상이 없을 경우 학습률을 0.3(factor)만큼 감소시키고 최소 학습률(min_lr)은 0.0001로 설정하였다.

본 논문에서는 StyleGAN 데이터 증강에 따른 흉부 X선 영상 및 복부 CT 영상 분류의 합성곱 신경망 분류 성능 및 효과를 정량적 및 정성적 평가를 통해 분석하였다. 첫째로 두 데이터셋에서의 StyleGAN 합성 영상을 관찰하여 StyleGAN 증강 기법의 효과 및 한계를 분석한다. 둘째, 신경망 분류결과에 대해 정확도, F1 score 및 클래스별 정밀도, 재현율 평가를 통한 정량적 성능 평가 및 비교를 수행한다. 셋째로 여러 혼합비율로 혼합된 증강 훈련 데이터로 학습된 분류기의 특징값을 tSNE 시각화를 통해 특징

Table 3: Performance comparison of chest X-ray pneumonia classification results.

Methods	Accuracy [%]	F1-Score [%]	Normal		Pneumonia	
			Precision [%]	Recall [%]	Precision [%]	Recall [%]
Baseline (StyleGAN 0%)	88.88±2.23	87.79±2.76	93.59±4.89	89.65±4.94	81.03±10.69	89.55±6.44
StyleGAN 25%	**91.83±0.34**	**91.21±0.34**	94.56±1.58	**92.56±1.12**	**87.26±2.27**	90.71±2.26
StyleGAN 50%	89.74±0.79	88.65±1.10	96.00±3.84	88.82±3.55	79.32±7.44	**93.09±5.29**
StyleGAN 75%	89.55±1.27	88.65±1.43	93.90±2.82	89.96±2.34	82.31±5.00	89.35±3.89
StyleGAN 100%	87.37±1.62	85.76±1.98	**96.56±0.77**	85.26±2.05	72.05±4.51	92.66±1.45

Table 4: Performance comparison of abdominal CT focal liver lesion classification results. (Sens.: Sensitivity, Spec.: Specificity)

Methods	Accuracy [%]	F1-Score [%]	Cyst		Hemgnaioma		Metastasis	
			Sens. [%]	Spec. [%]	Sens. [%]	Spec. [%]	Sens. [%]	Spec. [%]
Baseline (StyleGAN 0%)	72.25±1.01	60.71±1.12	85.63±2.11	**82.01±1.74**	29.33±1.49	89.46±1.95	69.40±1.22	84.94±0.69
StyleGAN 25%	**73.42±1.89**	**62.66±3.47**	**87.34±1.50**	80.67±1.29	33.33±11.79	91.55±3.35	69.53±3.31	85.06±2.59
StyleGAN 50%	73.16±1.73	61.97±1.76	85.78±2.17	77.88±1.40	28.00±5.58	94.58±2.81	71.41±2.91	82.41±1.04
StyleGAN 75%	71.14±0.99	59.63±1.38	81.41±1.69	78.10±1.45	24.67±8.69	**95.52±2.46**	**71.68±2.70**	76.58±1.61
StyleGAN 100%	40.98±0.87	35.75±9.04	56.88±24.33	64.47±36.08	**57.33±32.09**	62.45±21.16	24.03±13.56	**91.39±5.27**

값 분포도를 비교함으로써 실제 영상과 합성 영상의 역할과 혼합 비율에 따른 변화를 정성적으로 분석한다.

3.3 실험 결과

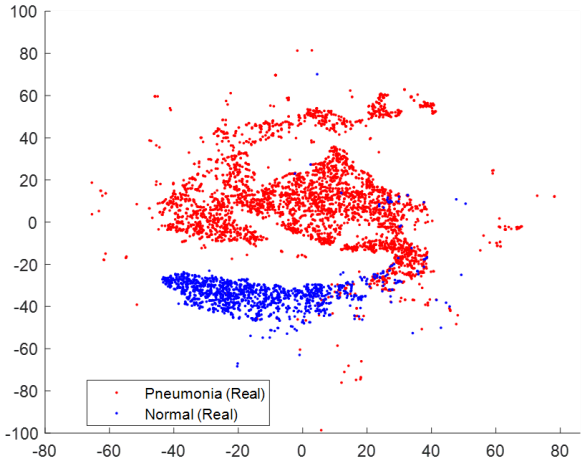
그림 4는 흉부 X선 영상에서 폐렴과 정상의 원본 영상과 StyleGAN을 통해 생성한 영상을 나타낸다. 그림 4 (a), (b)는 각각 폐렴과 정상 클래스의 원본 영상이고 그림 4 (c), (d)는 각각 폐렴과 정상 클래스의 StyleGAN 합성 영상이다. 그림 4 (a), (b)와 그림 4 (c), (d)는 육안으로 봤을 때 거의 유사한 것을 볼 수 있다. 어파인 변환 기반 데이터 증강 기법은 기존 영상을 변형하는데 그치는데 반해 StyleGAN으로부터 생성한 영상들은 새로운 조합이 가능하다. 하지만 그림 4 (c)의 2, 3번째 영상과 그림 4 (d)의 2, 3번째 영상처럼 StyleGAN이 영상을 생성하는 데 있어서 한정된 패턴이 계속해서 나오는 모드 붕괴 현상이 발생한다는 한계가 있다.

그림 5는 복부 CT 영상에서 낭종, 혈관종, 전이암의 원본 영상과 StyleGAN 생성 영상을 나타낸다. 흉부 X선 영상에서와 마찬가지로 StyleGAN이 실제영상과 유사한 외양과 각 클래스 별 특징을 가지는 합성영상을 생성하는 것을 확인할 수 있다.

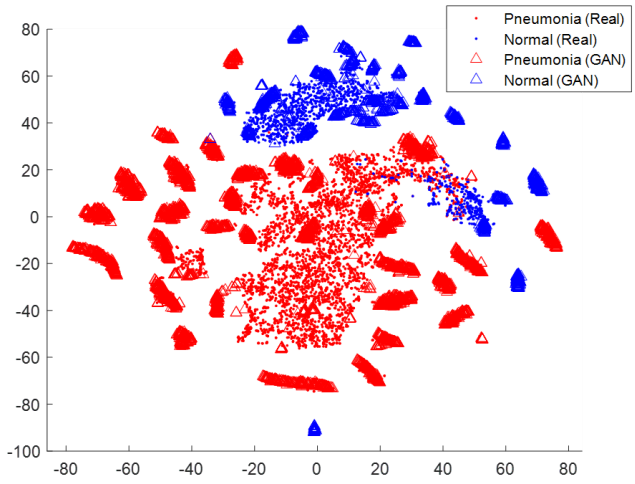
표 3는 흉부 X선 영상에서 훈련 데이터의 StyleGAN 혼합 비율에 따른 폐렴 분류결과의 정량적 평가를 나타낸다. 합성영상만으로 훈련한 데이터(StyleGAN 100%)를 제외하고는 StyleGAN 합성 영상을 함께 훈련하는 것이 분류 정확도를 모두 개선하는 것으로 나타났다. 특히, 평균 정확도와 f1 score, 정상의 재현율과 폐렴의 정밀도는 25% 포함 데이터에서 가장 높은 수치의 결과를 보였고, 정상의 정밀도는 100% 포함 데이터에서, 폐렴의 재현율은 50% 포함 데이터에서 가장 높은 수치의 결과를 보였다. 따라서 25% 포함 데이터의 성능이 가장 좋은 것으로 분석되는데, 평균 정확도는 91.83%로 어파인 변환 기반 데이터 증강에 비해 2.95% 높게 나타났고, f1 score는 91.21%로 어파인 변환 데이터 증강에 비해 3.42% 높게 나타났다. 정상의 정밀도와 재현율, 폐렴의 정밀도와 재현율 또한 어파인 기반 데이터 증강에 비해 각각 0.97%, 2.91%, 6.23%, 1.16% 높은 수치의 결과를 보였다. 실제 영상만을 사용하는 것보다는 StyleGAN 합성 영상을 25%의 비율로 함께 사용하는 것이 전반적인 성능을 개선하는 효과를 보였다.

표 4은 복부 CT 영상에서 훈련 데이터의 StyleGAN 혼합 비율에 따른 간전이암 분류결과의 정량적 평가를 나타낸다. 기본 Baseline 모델의 결과를 보면 전체 질병의 특이도와 낭종의 민감도는 80% 대로 나타나는 반면 전이암의 민감도는 69.4%, 특히 소수 클래스인 혈관종의 민감도는 29.33%로 매우 낮게 나타나는 것을 확인할 수 있다. 이는 클래스 불균형 등으로 인해 소수 클래스인 혈관종의 학습이 매우 어려운 조건임을 알 수 있다. 반면 StyleGAN 25% 포함 데이터에 의해 훈련된 모델은 평균 정확도 73.42%, 평균 F1 score 62.66%로 주어진 과제에서 가장 높은 분류 정확도를 달성하였으며 StyleGAN 합성 영상을 포함하지 않았을 때보다 정확도는 약 1.17%, F1 score은 약 1.95% 상승된 수치로 분류 성능이 향상되었음을 알 수 있다. 클래스 별 성능을 관찰하면 낭종의 경우 StyleGAN 25%에서 가장 개선된 민감도를 보이는 반면 전이암은 StyleGAN 75%에서 가장 민감도가 개선되는 결과를 보였다. 혈관종은 StyleGAN 100%에서 가장 높은 민감도를 보였으나 이는 낭종 및 전이암의 분류 성능이 떨어진 데에 대한 반사이익으로 보이며 이를 제외하고는 StyleGAN 25%에서 가장 개선된 민감도를 보였다.

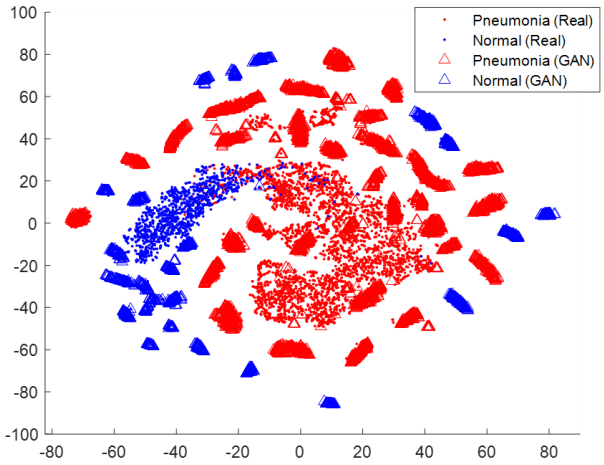
훈련 데이터셋에서 실제영상과 합성영상의 혼합 비율에 따라 이처럼 분류 결과가 차이나는 이유를 분석하기 위해 각 분류기의 특징값 분포를 tSNE 시각화를 통해 분석하였다. 그림 6은 흉부 X선 영상의 폐렴 분류에서 원본 영상과 서로 다른 비율로 혼합된 StyleGAN 합성 영상에 의해 훈련된 분류기를 통해 추출된 특징들의 분포를 t-SNE 기법으로 시각화한 결과이다. 그림 6 (a)에서 원본 영상 데이터의 분포를 보면, 폐렴과 정상 데이터가 각각 큰 클러스터를 형성하면서도 경계영역에서 혼동되는 데이터



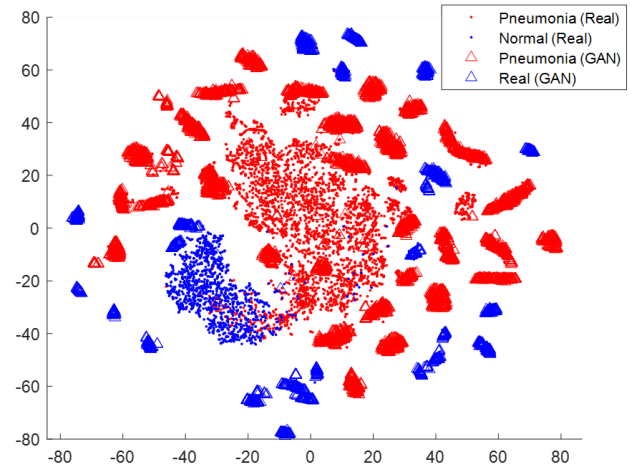
(a) Baseline (StyleGAN 0%)



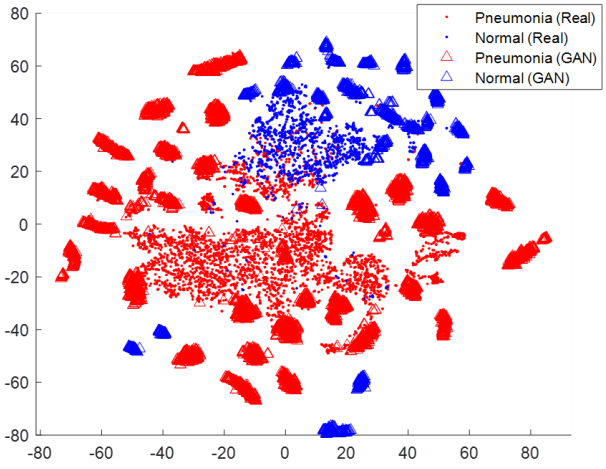
(b) StyleGAN 25%



(c) StyleGAN 50%

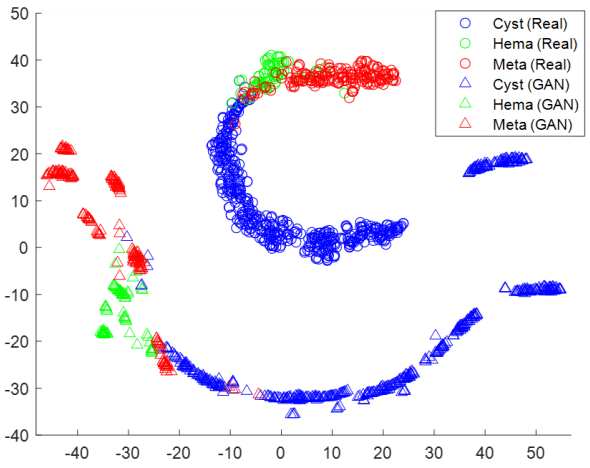


(d) StyleGAN 75%

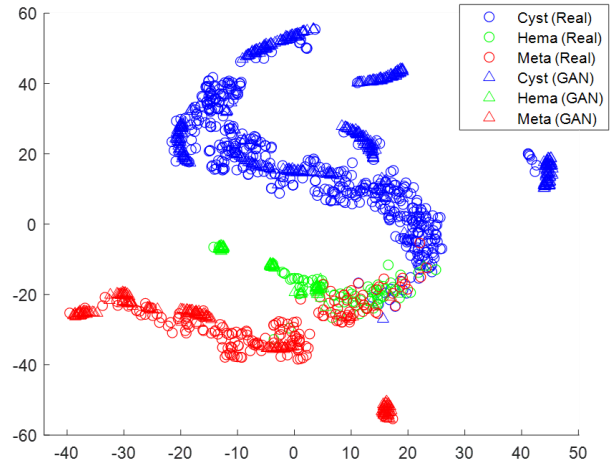


(e) StyleGAN 100%

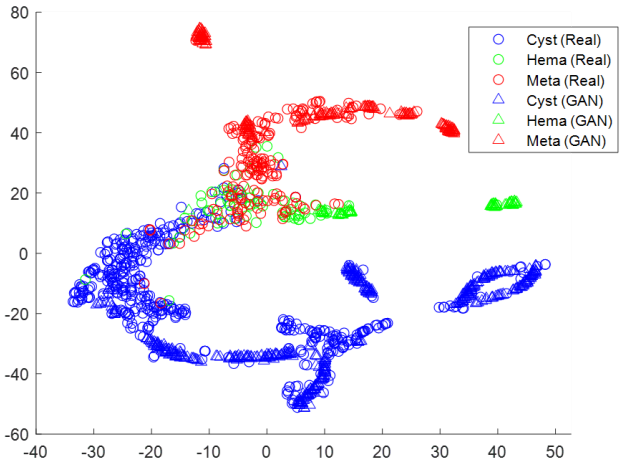
Figure 6: Comparison of the t-SNE distributions of real and synthetic image features in various chest X-ray classification networks. Each percentage represents the proportion of StyleGAN synthetic images in the training dataset.(Normal: blue, Pneumonia: red, Real images: dots, StyleGAN images: triangles)



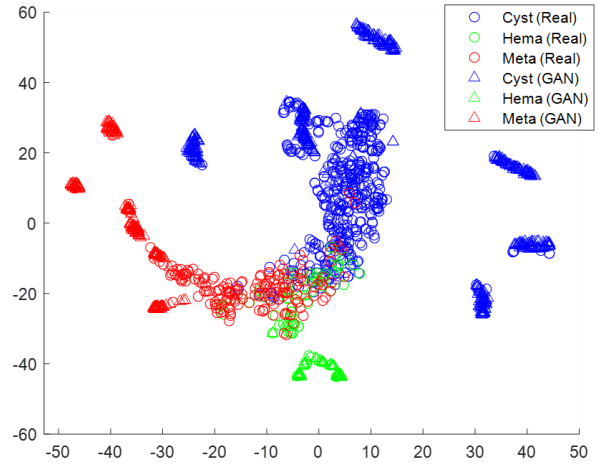
(a) Baseline (StyleGAN 0%)



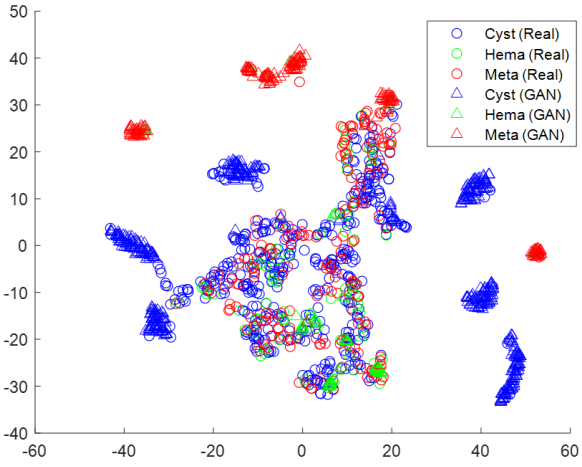
(b) StyleGAN 25%



(c) StyleGAN 50%



(d) StyleGAN 75%



(e) StyleGAN 100%

Figure 7: Comparison of the t-SNE distributions of real and synthetic image features in various CT focal liver lesion classification networks. Each percentage represents the proportion of StyleGAN synthetic images in the training dataset. (Cyst: blue, Hemangioma: green, Metastasis: red, Real images: circles, StyleGAN images: triangles)

들이 관찰된다. 정상 데이터는 대부분 한 클러스터에 분포하나, 폐렴 데이터는 좌측 정상 데이터 부분을 포함해 우측 큰 클러스터의 외곽까지 분산되어 있다. 이러한 분포 차이는 폐렴 환자의 정밀도가 정상에 비해 상대적으로 낮은 경향을 보이는 현상을 설명할 수 있다. 그림 6 (b)는 StyleGAN 합성 영상이 25% 포함된 데이터에서 훈련된 분류기에서 추출한 원본 영상 데이터와 GAN 영상 데이터의 t-SNE 특징값 분포를 나타낸다. 원형으로 나타난 StyleGAN 데이터의 특징은 두 가지로 요약할 수 있다. 첫째, StyleGAN 데이터는 각 클래스 분포의 내부 또는 클래스 간의 경계영역보다 클래스의 테두리 영역에 집중적으로 분포한 것을 확인할 수 있다. 이는 StyleGAN 데이터가 클래스의 평균 이미지 또는 클래스 간의 경계영역에 있을 법한 모호한 영상을 생성하기보다 각 클래스의 특징이 분명한 영상을 집중적으로 생성함을 시사한다. 둘째, StyleGAN 데이터는 클래스 분포 전역에 흩뿌려져 있지 않고 군데군데 집중적으로 몰려 있는 것을 확인할 수 있다. 이는 특정 패턴의 영상들이 반복해서 생성되는 모드 붕괴(mode collapse) 현상이 여전히 나타남을 의미한다. StyleGAN 25%과 같이 실제 영상의 비율이 더 클 때에는 클래스 분포가 유지되면서 StyleGAN의 확장이 이를 보완하는 역할을 수행하지만, StyleGAN 50%와 StyleGAN 75%에서는 클래스 내부를 유지하고 클래스 간 경계를 정의하는 실제 영상의 역할이 줄어들면서 StyleGAN 합성 영상의 분포가 실제 영상의 클래스와 분리되기 시작하는 것을 볼 수 있다. 실제 영상 없이 StyleGAN 합성 영상만으로 학습한 StyleGAN 100%의 경우 하나의 구심점이 되는 클래스 분포 없이 합성 영상의 분포가 모드 붕괴의 패턴 별로 따로 떨어져 산개해 있는 것을 확인할 수 있다. 이를 통해 특징값 공간(feature space)에서의 StyleGAN 합성 영상의 특징은 클래스 특징이 두드러지는 데이터를 생성함으로써 클래스 외곽으로 분포를 확장하고 강화하는 반면, 클래스의 내부 분포를 채워서 하나의 클래스 분포를 완성하고 클래스 간의 경계를 구분하는 역할은 실제 영상들이 하게 됨을 알 수 있다. 따라서 각 클래스의 외곽을 보강하고 강화하는 StyleGAN 영상과 이들 합성 영상들을 하나의 분포로 모으고 클래스 간의 경계를 정의해줄 수 있는 실제 영상 사이의 적절한 비율로 훈련데이터를 구성하고 학습할 때 분류 효율을 개선할 수 있다는 것을 확인하였다.

그림 7은 복부 CT 영상의 간전이암 분류에서 원본 영상과 서로 다른 비율로 혼합된 StyleGAN 합성 영상에 의해 훈련된 분류기를 통해 추출된 특징들의 분포를 t-SNE 기법으로 시각화한 결과이다. 그림 7 (a)에서 원본 영상 데이터의 분포를 보면 낭종과 전이암 클래스가 큰 클러스터를 형성하면서 두 클래스의 경계 부근에 혈관종 클래스가 겹쳐져서 세 클래스가 혼동되는 영역이 관찰된다. 이러한 분포 특성은 낭종과 전이암의 민감도가 비교적 높은 반면 혈관종의 민감도가 낮은 경향을 보이는 정량적 결과를 설명할 수 있다. 그림 7 (b)에서 나타나는 StyleGAN 합성 영상은 첫째, 클래스 내부 또는 클래스 간의 경계보다 클래스의 외곽 영역에 집중적으로 분포하는 확장 특성을 보이며, 둘째, 클래스 전체에 걸쳐 데이터가 산개해 있기 보다 특정 패턴을 중심으로 데이터들이 뭉쳐서 분포하는 모드 붕괴 현상을 보이는 점에서 앞의 흉부 X선 영상 분류에서의 분석과 일치한다. 마찬가지로 이러한 특성으로 인해 합성 영상과 실제 영상의 비율에 따른 클래스의 분포 변화도 설명되는데, StyleGAN 25%와 StyleGAN 50%의 경우에는 실제 영상의 클래스 분포를 중심으로 합성 영상이 외곽을 확장하는 방식으로 분포하는 반면, StyleGAN 75%에서는 클래스 내부 분포를 유지하는 실제 영상의 양이 감소하여 합성 영상들이 실제 영상의 분포와 분리되어 나타나고, StyleGAN 100%에서는 구심점이 되는 클래스가 없이 각 합성 영상의 모드 붕괴 패턴에 따라 산개한 방식으로 분포하게 된다. 이를 통해 흉부 X선 영상에서와 마찬가지로 복부 CT 영상 분류에서도 합성 영상의 확장 특성과 실제 영상의 클래스의 구심이 되고 클래스 간 경계를 정의하는 역할 사이의 적절한 비율을 통해 분류 효과를 개선할 수 있음을 확인할 수 있다.

4. 결론

본 연구에서는 의료영상 분류를 위한 합성곱 신경망 학습에서의 StyleGAN 합성 영상의 데이터 증강 효과를 평가 및 분석하였다. 이를 위해 흉부 X선 영상에서의 폐렴 진단과 복부 CT 영상에서의 간전이암 분류의 두 가지 의료영상 분류 문제에 대해 StyleGAN 데이터 증강 및 신경망 학습을 적용하였다. 실험에서는 StyleGAN 합성 영상에 대한 육안 평가와 훈련데이터에서 합성 영상 및 실제 영상의 비율에 따른 정확도, F1-score 등의 정량적 평가 및 tSNE를 통한 특징값 분포 관찰 및 비교를 수행하였다. 합성 영상의 육안 평가에서는 StyleGAN이 실제 영상의 외양과 각 클래스별 특징을 살린 합성 영상을 생성해내는 것을 확인하였으며 특정 패턴의 영상들이 반복적으로 생성되는 모드 붕괴현상이 일어남을 확인하였다. 정량적 평가에서는 두 가지 분류 문제 모두 StyleGAN 합성 영상이 25% 포함된 데이터에서 훈련된 분류기가 가장 개선된 정확도와 F1-score를 보였다. 클래스 별 성능에 있어서도 StyleGAN 합성 영상이 추가된 학습에서 각 클래스의 민감도가 개선되는 결과를 보였다. tSNE 분포도 관찰을 통한 정성적 분석에서는 StyleGAN 합성 영상이 클래스의 내부나 클래스 간 경계보다는 클래스의 특징이 뚜렷한 외곽 영역의 데이터를 보강함으로써 클래스를 외곽으로 확장하는 역할을 수행하며, 클래스 전체에 걸쳐 고루 분포하기보다는 특정 패턴들에 집중되어 나타나는 모드 붕괴 특성을 보임을 두 분류 문제에서 일관되게 확인하였다. 더불어 클래스 외곽을 확장하는 StyleGAN 영상의 특성과 클래스 내부와 클래스 간 경계를 정의하는 실제 영상이 적절한 비율로 혼합됐을 때 분포 특성 및 분류 성능이 가장 개선되고, StyleGAN 영상의 비율이 높아질수록 합성 영상 분포와 실제 클래스 분포 간의 분리가 일어나면서 분류 성능이 떨어지는 결과가 관찰됨을 확인하였다. 향후 연구에서는 MixUp, CutMix와 같이 클래스 간 경계 영역을 강화하는 것으로 알려진 데이터 증강 기법을 통해 StyleGAN 증강 기법의 효과를 보완하여 분류기 성능을 추가적으로 개선할 수 있다.

감사의 글

이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. RS-2023-00207947)이며, 서울여자대학교 학술연구비의 지원에 의한 것임(2024-0220). 본 논문에서 사용한 복부 CT 영상 데이터를 제공해주신 세브란스병원 영상의학과 임준석 교수님께 감사의 말씀을 드립니다.

References

[1] J. Garstka and M. Strzelecki, "Pneumonia detection in x-ray chest images based on convolutional neural networks and data augmentation methods," in *2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications*. Institute of Electrical and Electronics Engineers, 2020, pp. 18–23.

[2] S. Motamed, P. Rogalla, and F. Khalvati, "Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images," *Informatics in Medicine Unlocked*, vol. 27, p. 100779, 2021.

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations*, 2016.

[4] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2017.

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Computer Vision and Pattern Recognition*, 2019.

[6] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks," in *Institute of Electrical and Electronics Engineer international conference on acoustics, speech and signal processing*, 2018.

[7] 김덕선, 우아라, 이한상, and 홍헬렌, "복부 ct 영상에서 심층 합성곱 신경망 기반의 국소 간 병변 분류를 위한 데이터 증강 기법의 효과," *한국컴퓨터그래픽스학회논문지*, vol. 29, no. 2, pp. 1–11, 2023.

[8] D. Zhao, D. Zhu, J. Lu, Y. Luo, and G. Zhang, "Synthetic medical images using f&bgan for improved lung nodules classification by multi-scale vgg-16," *Symmetry*, vol. 10, no. 10, p. 519, 2018.

[9] H. Lee, H. Lee, H. Hong, H. Bae, J. Lim, and J. Kim, "Classification of focal liver lesions in ct images using convolutional neural networks with lesion information augmented patches and synthetic data augmentation," *Medical Physics*, vol. 48, no. 9, pp. 5029–5046, 2021.

[10] A. Woo, H. Lee, J. Lim, and H. Hong, "Classification of focal liver lesions in abdominal ct images using convolutional neural networks with stylegan data augmentation," in *Proceeding of the Fall Conference of the Korea Multimedia Society*, vol. 26, no. 1, 2023, pp. 94–96.

[11] M. Loey, S. Florentin, and N. Khalifa, "Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning," *Symmetry*, vol. 12, no. 4, p. 651, 2020.

[12] S. Buragadda, K. Rani, S. Vasantha, and M. Chakravarthi, "Hcugan: Hybrid cyclic unet gan for generating augmented synthetic images of chest x-ray images for multi classification of lung diseases," *International Journal of Engineering Trends and Technology*, vol. 70, no. 2, pp. 229–238, 2022.

[13] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[14] R. Hao, K. Namdar, L. Liu, M. Haider, and F. Khalvati, "A comprehensive study of data augmentation strategies for prostate cancer detection in diffusion-weighted mri using convolutional neural networks," *Journal of Digital Imaging*, pp. 862–876, 2021.

[15] M. Kim and H. Bae, "Data augmentation techniques for deep learning based medical image analyses," *Journal of the Korean Society of Radiology*, vol. 81, no. 6, 2020.

[16] M. Nishio, S. Noguchi, H. Matsuo, and T. Murakami, "Automatic classification between covid-19 pneumonia, non-covid-19 pneumonia, and the healthy on chest x-ray image: combination of data augmentation methods," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[17] D. Kermany, M. Goldbaum, W. Cai, C. Valentim, H. Liang, S. Baxter, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[18] K. Asnaoui, Y. Chawki, and A. Idri, “Automated methods for detection and classification pneumonia based on x-ray images using deep learning,” in *Artificial intelligence and blockchain for future cybersecurity applications*. Cham: Springer International Publishing, 2021, pp. 257–284.

[19] I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell, “Epidemiology and etiology of childhood pneumonia,” *Bulletin of the World Health Organization*, vol. 86, pp. 408–416B, 2008.

[20] H. Zar, S. Andronikou, and P. Nicol, “Advances in the diagnosis of pneumonia in children,” *British Medical Journal*, vol. 358, 2017.

〈 저 자 소 개 〉

이 한 상
- 2011년 2월 한국과학기술원 전기 및 전자공학과 졸업(학사)
- 2013년 2월 한국과학기술원 전기 및 전자공학과 졸업(석사)
- 2019년 2월 한국과학기술원 전기 및 전자공학과 졸업(박사)
- 2019년 3월~현재 한국과학기술원 정보전자연구소 연수연구원 재직 중
- 관심분야 : 인공지능, 딥러닝, 컴퓨터 비전, 의료영상처리 및 분석
- https://orcid.org/0000-0003-0917-1589

우 아 라
- 2024년 2월 서울여자대학교 소프트웨어융합학과 졸업(학사)
- 관심분야 : 인공지능, 딥러닝, 기계학습, 생성모델
- https://orcid.org/0009-0003-9299-2237

홍 헬 렌
- 1994년 2월 이화여자대학교 전자계산학과 졸업(학사)
- 1996년 2월 이화여자대학교 전자계산학과 졸업(석사)
- 2001년 8월 이화여자대학교 컴퓨터학과 졸업(박사)
- 2001년 9월~2003년 7월 서울대학교 컴퓨터공학부 BK 조교수
- 2006년 3월~현재 서울여자대학교 소프트웨어융합학과 교수
- 관심분야 : 의료 인공지능, 딥러닝, 영상처리 및 분석
- https://orcid.org/0000-0001-5044-7909